# An a contrario space-time grouping framework for the detection of coherent motions

Thomas Veit, Frédéric Cao, Patrick Bouthemy

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *An* a contrario *space-time grouping framework for the detection of coherent motions*

Thomas Veit — Frédéric Cao — Patrick Bouthemy

**N° 6061**

Novembre 2006

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6061--FR+ENG

# An *a contrario* space-time grouping framework for the detection of coherent motions

Thomas Veit , Frédéric Cao , Patrick Bouthemy

Thème COG — Systèmes cognitifs
Projet Vista

Rapport de recherche n° 6061 — Novembre 2006 — 33 pages

**Abstract:** This paper presents a method for detecting independent temporally-persistent motion patterns in image sequences. The result is a description of the dynamic content of video sequences in terms of moving objects, their number, image postion and approximate motion. It provides for each detected motion pattern a local trajectory as well as a confidence level in the detection. The method is based on local motion measurements extracted from short video segments. These measurements are mapped in an adequate grouping space where independent trajectories correspond to distinct clusters. The automatic cluster detection is handled in an *a contrario* framework, which is general and involves no parameter tuning. The method was successfully applied to real video sequences featuring rigid and non-rigid moving objects, static and mobile cameras, and distracting motions. The output of this method could initialize tracking algorithms. Applications of interest are robot navigation, car-driver assistance, video surveillance and activity recognition.

**Key-words:** coherent motion detection, local trajectories, *a contrario* grouping, visual motion analysis

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

# Détection de mouvements cohérents par groupement spatio-temporel *a contrario*

**Résumé :** Ce document présente une méthode pour détecter des motifs de mouvements indépendants persistants au cours du temps. Cette méthode permet d'obtenir une description du contenu dynamique d'une séquence vidéo en termes d'objets mobiles : leur nombre, leurs positions dans l'image et leurs déplacements. Chaque motif de mouvement détecté est caractérisé par une trajectoire locale et un niveau de confiance. La méthode s'appuie sur l'accumulation de mesures locales de déplacement sur des segments vidéo courts. Dans un espace de groupement soigneusement choisi les trajectoires indépendantes correspondent à des groupes distincts de mesures. Un algorithme de détection *a contrario* permet d'extraire ces groupes automatiquement. La méthode a été testée avec succès sur des séquences vidéo réelles aux contenus variés : objets mobiles rigides et non-rigides, caméra statique ou mobile, présence de mouvements parasites. Les éléments de trajectoires extraits par cette méthode peuvent servir à initialiser de manière robuste des algorithmes de suivi. Les applications possibles sont la navigation en robotique, l'assistance à la conduite, la vidéo-surveillance ainsi que la reconnaissance de contenus.

**Mots-clés :** détection de mouvements cohérents, trajectoires locales, groupement *a contratrio*, analyse du mouvement visuel

# 1 Introduction

## 1.1 Problem setting

A general problem in motion analysis is the early reliable detection of pieces of trajectories of moving objects in natural image sequences. Accurately and efficiently solving this problem is of crucial interest for applications such as robot navigation and car-driver assistance (involving mobile obstacle detection and avoidance), or video-surveillance and human activity recognition. According to Ullman [1], the most fundamental questions when analysing the dynamic content of a video sequence are (in increasing order of complexity):

1. Are there moving objects in the observed scene?
2. How many?
3. Where are they?
4. What is their motion?

The method proposed in this paper aims at answering these four questions within a unified framework. The overall objective is to detect temporally-persistent independent motion patterns. In other words, the goal is to detect one short-term trajectory for each moving object of the scene. Based on characteristic image features, local motion measurements are extracted from the image sequence and mapped into a well-specified motion space. In this grouping space, independent objects moving along trajectories form clusters. These clusters are detected automatically by means of an innovative *a contrario* cluster detection framework. The involved cluster detection algorithm is fully automatic and provides a confidence level for each detected object trajectory.

It seems to us that there is a gap to be filled between two types of issues. On one hand, there are motion detection methods. Most methods are actually closer to change detection, since they make decision on very local time intervals, with no real search of any spatio-temporal coherence [2, 3]. As a consequence, significant moving objects cannot be distinguished from "parasitical" motion. The temporal content alone is usually very noisy; hence, local spatial (and possibly temporal) regularity is usually introduced, which is the simplest mean to enforce temporal coherence [4]. On the other hand, if the position of a given moving object is known, efficient methods allow one to track them. Many algorithms are variations or extensions of the celebrated Kalman filter. Recent progress based on the non-linear particle filtering approach led to very impressive results able to handle occlusions, shape deformation, etc [5, 6, 7]. The weak point of these methods is their initialization which is usually supervised.

The method proposed in this paper may be considered as addressing simultaneously coherent motion detection and track initialization. The purpose is to decide upon the existence of small pieces of trajectories on short durations (typically 10 or 20 frames). Detection thresholds for extracting these pieces of trajectories are computed automatically. It is clear that such thresholds exist also from a perceptual point of view. As an example, a slowly moving object has to be observed for a long time to be detected. Hence, there should be a relation between the size of an object, its velocity, the duration of observation and its detectability. When dealing with digital image sequences, detectability is also influenced by image quality. The method described in this paper uses a detection principle, intuited by Helmholtz and formulated by Desolneux, Moisan and Morel [8] (also following works by Attneave [9] and Lowe [10]). It states that a particular configuration is perceptually relevant if it cannot occur by chance, i.e., it contradicts a general random structure of the observations.

## 1.2 Overall strategy

The purpose of this work is to extract geometrical evidence for moving objects from a set of successive digital images (about 10-20). More precisely, is it possible to prove that image parts along a sequence display locally a coherent motion, and define a piece of trajectory? With which degree of confidence?

The strategy is the following. First, local motion measurements are extracted from successive pairs of images. These measurements are based on characteristic image features such as similarity invariant pieces of level lines [11], SIFT descriptors [12] or KLT features [13, 14]. These features have to be local enough, because of partial occlusions, shadows, etc. If the duration of observation is short enough, the motion of objects is approximately rectilinear with a constant velocity. This velocity, as well as the position of the shape element at time $t = 0$ is, in this simple case, completely determined by the displacement between two images. This results in a point in $\mathbb{R}^4$: two real coordinates for the velocity and two for the initial position. Now, if these pairs correspond to the same moving object in different frames, then the corresponding points form clusters in $\mathbb{R}^4$. As a consequence, the detection of pieces of trajectories results in a cluster detection problem.

Let us consider $M$ data points, $X_1$, ..., $X_M$ in $\mathbb{R}^4$, each corresponding to a couple (initial position,velocity), possibly detected at different instants. Following the same argument as in [15], an *a contrario* method is adopted: assume all the pairs are casual, and do not correspond to a coherent trajectory. Then, it is sound to assume that the $X_i$ are independent and identically distributed according to a probability distribution to be specified. It is very unlikely that an important proportion of the $X_i$'s can be observed in a single small region of $\mathbb{R}^4$. Whenever this is actually observed, then the hypothesis that the $X_i$ are random is certainly false, and some of them should be grouped. Natural questions arise, that are answered in this paper: how many groups are there (if any)? Which groups are relevant? Is it possible to quantify the meaningfulness of a group of points? How to select among nested groups?

The outline of the paper is the following. Section 2 presents some related work. Section 3 describes how to extract local motion measurements based on image features and how to map them in an adequate motion grouping space. Section 4 introduces the *a contrario* grouping method and details its application to the detection of coherent motions. Section 5 experimentally validates the theory. Conclusion and perspectives are given in Section 6.

## 2 Related work

Different approaches to exhibit temporal motion coherence in image sequences have been developed. A first group of methods, attempts to directly analyze the characteristics of motion over time or to extract some structures from the space-time volume defined by an image sequence. A second class of methods addresses the detection of coherent motions as a grouping problem. Most of these methods lack an efficient clustering framework. Finally, our method shares some ingredients with Structure From Motion methods, namely, the use of image features and clustering algorithms.

In [16], Wixson proposes to accumulate directionally consistent optical flow. An estimate of the total image distance moved by each pixel during the sequence enables to discriminate between objects moving with a consistent direction and parasitical motion. Gryn et al. [17] have specified even more precise motion templates, driven by the application, in other words trading generality for better computational efficiency. Different methods attempt to analyze the space-time volume of image sequences. For instance, Ricquebourg and Bouthemy [18] as well as Sarkar et al. [19] look for motion structures (typically alignments) in spatio-temporal slices. The same type of idea is used by Kornprobst and Medioni [20] where trajectories are the result of a vote. Another approach to coherent motion detection developed by Laptev et al. [21] is to exploit space-time interest points. Focusing on the class of periodic motions enables for example to extract pedestrians in cluttered environments. One of the most difficult issues in that context is the automatic computation of robust detection thresholds.

If local motion measurements are suitably parametrized, the detection of independent coherent motions can be viewed as a clustering problem. Yuille and Grzywacz[22] proposed a clustering approach after suitably representing visual patterns, and attempted to classify the typical configurations of visual motion. A complex observation would be a combination of these elementary motion templates, that should be detected by a grouping procedure. However, their work remains formal with no computational theory. Burgi et al.[23] propose a Bayesian framework along with a generative model of trajectory. More recently, Gao et al.[24] worked on motion detection *via* clustering. Motion information is extracted using edge elements which are grouped according to spatial proximity and motion persistence over time. The clustering strategy relies on several user-set parameters. This certainly harms the generality of the method.

The similarity of the ingredients involved in our method with those involved in Structure From Motion (SFM) methods might be misleading. The focus of SFM methods is more on characterizing the 3D geometry of the scene than on detecting coherent motion patterns [25, 26]. The presence of one or several moving objects is assumed and therefore the detection issue is not addressed. Furthermore, the features detected in the image sequences need to be tracked through all the sequence [27, 28]. This requirement is obviously difficult to meet in the presence of occlusions or noisy image sequences. Factorization methods usually rely on spectral clustering for the clustering step. This clustering method, based on algebraic matrix manipulations, is known to be very sensitive to noise. Other methods rely on iterative optimization methods to build clusters, for example Expectation-Maximisation or K-means [29]. These methods require the number of clusters to be specified. Moreover, the results are sensitive to initialization. An alternative is to resort to model selection to determine the number of moving objects. In [30], a rank constraint is developed to estimate the number of moving objects. Torr and Murray [31] propose a stochastic clustering method to group local motion measurements from several moving objects based on 3D geometry. They address the different issues of clustering, namely cluster validity assessment and merging of clusters. Their method relies on the combination of several heterogeneous criteria involving several parameters. Their method is based on two frames and the clustering is therefore rather based on shape than on motion coherence over time.

# 3 Image features and local displacements measurements

The features to be extracted from images must be local (because of possible partial occlusions), stable, and invariant enough to the deformations an object may encounter through a sequence (approximate rigid motion, contrast change...). Different type of features meet these requirements:

- Similarity Invariant Pieces of Level Lines (SIPLL) [11],
- SIFT descriptors [12],
- KLT features [13].

The reader is referred to these articles for the exact definition and the computation of these features. Each type of features has its advantages and drawbacks. The three types of features tested differ in terms of invariance to geometrical transformations, discriminative power and computational load. The first type of features is a local piece of contrasted level lines (isophotes), as detailed in [32]. The main advantage is that its associated representation is invariant with respect to contrast change and similarity transformations. When the image resolution is fine enough, this first type of features is accurate since level lines locally coincide with edges. On the other hand, the computational load is a bit heavy. Besides, satisfying the largest invariance group is useful when attempting to match images if there is no *a priori* knowledge that they have some content in common. When matching two consecutive images in a video, requiring such a degree of invariance may be unnecessary. The second type of features are SIFT descriptors [12]. They are slightly less invariant than SIPLL, and less intuitive from a geometric point of view but faster to compute. They have proved very efficient for matching multiple views of a single scene. Still in decreasing order of complexity and invariance are KLT features [13, 33] obtained by correlation of patches around interest points (Harris points [34] in the original version). In contrast with SIPLL and SIFT descriptors, the KLT extraction framework includes the computation of a local displacement vector. Let us point out that our detection method is independent of the type of features and could therefore easily adapt to other type of features.

Given a pair of successive images of the sequence at time instants $t$ and $t+1$, any of these features enables to compute local motion measurements. In the case of SIPLL or

SIFT descriptors, a displacement measurement is obtained by matching a feature in the first frame with its best corresponding feature in the next frame. Of course, when looking for a match, the whole image does not need to be explored. Since object displacements in the image are limited (typically less than 10 pixels between two consecutive frames), focusing on a neighborhood of the feature position in the first image is sufficient. For example, it is reasonable to restrict the matching process to features in the second frame within a distance of 20 pixels from the position of the feature in the first frame. Now, the difference between the position $x_t$ at time instant $t$ and $x_{t+1}$ at time instant $t+1$ provides the displacement $v$. For KLT features, the displacement $v$ is directly computed by an optimization process involving both image frames [13]. Let us define the vector $(x^{\text{ref}}, v) \in \mathbb{R}^4$ by $x^{\text{ref}} = x_t - t\ v$. By first order approximation, the velocity $v$ is constant and $x^{\text{ref}}$ would be the theoretical initial position of the feature at time instant $t = 0$. This hypothesis is sound if the duration of observation is short enough. Moreover, let us point out that the aim is not to measure accurately the characteristics of motion, but only to robustly detect pieces of trajectories. Hence, this hypothesis does not need to be satisfied very accurately.

Now, a part of the same moving object at different time instants, or different parts of the same moving object should lead to approximately the same values of initial position and velocity. Therefore, local motion measurements are accumulated over several successive pairs of frames. The total number of frames should be large enough so that clusters contain a sufficient number of data points in order to be detected. The total observation time should remain low so that the first order approximation on the trajectory remains valid. Typically, the number of frames involved ranges from 3 to 30. Let us emphasize that a given feature does not need to be tracked through all the frames. This makes the proposed method robust to noise, appearance changes, as well as partial and global occlusions. Fig. 1 schematically describes how local displacement measurements corresponding to objects following trajectories lead to clusters in the four-dimensional grouping space $(x^{\text{ref}}, v)$. Local motion measurements in the images corresponding to the same trajectory accumulate and form clusters in the grouping space $(x^{\text{ref}}, v)$. Fig. 2 displays the two-dimensional projections of the couples $(x^{\text{ref}}, v) \in \mathbb{R}^4$ extracted from 10 successive frames of a highway surveillance sequence. The middle plot corresponds to $x^{\text{ref}}$, i.e., the vertical coordinates vs. the horizontal coordinates of the theoretical initial position. The right plot corresponds to the polar coordinates of $v$, orientation vs. magnitude. Three clusters in $\mathbb{R}^4$ can be distinguished corresponding to the three moving objects that appear in the scene displayed in the left image. Automatically detecting clusters in this four-dimensional grouping space results in detecting the independent motion patterns that are temporally coherent, in other words the three moving objects. Local motion measurements corresponding to the background of the scene are scattered in position and velocity direction but highly concentrated at velocity magnitude 0. They do not form a distinct cluster in $\mathbb{R}^4$.

In order to deal with mobile cameras, dominant motion estimation and motion compensation are applied. A general and robust dominant motion estimation algorithm is applied [35]. The dominant motion is identified with camera motion. This identification is possible under some hypotheses such as the image size of the moving objects and the absence of significant depth discontinuities in the background. These hypotheses are usually verified in typical surveillance videos. Once the camera motion is compensated, local motion measurements corresponding to the background display almost null velocity exactly as in the static camera case.

Since the computational load of the grouping procedure directly depends on the number of local motion measurements, discarding local motion measurements that obviously belong to the background dramatically saves computation time. Two simple strategies to discard background measurements can be adopted. If for each image of the sequence a detection map is available that indicates which regions of the image belong to the background and which regions are moving, only features corresponding to moving regions can be processed. For example, such a detection map can be obtained by applying an automatic moving region detection as proposed in [36]. This strategy is preferred when working with SIPLL or SIFT descriptors. The other strategy consists in discarding all features with an estimated inter-frame velocity magnitude smaller than a given threshold, typically 1 pixel. This threshold corresponds to the image sampling rate and is not very demanding. This second strategy is preferred when working with KLT features. Features remaining after discarding those belonging to the background are termed *moving features*. When applied to *moving features*, the task of the clustering procedure is to detect groups of features corresponding to each object moving independently and consistently over time. A similar background subtraction strategy is adopted in [28].

# 4 Coherent motion detection by *a contrario* clustering

This section presents an efficient clustering algorithm that enables to answer the questions of Section 1.1 in a unified framework. Let us consider a set of points $\{X_1, ..., X_M\}$ in $\mathbb{R}^4$. Does this set contain any group? How many, and how meaningful are they? This problem is one of the numerous forms of cluster analysis. While many classical efficient techniques [37] propose sound cluster candidates, the above questions do not have a definitive answer. In particular, it is difficult to make a robust decision about the existence of a group (known as the problem of *validity*), or whether it should be cut into subgroups or not. This is precisely the problems this section deals with. Some ideas presented here have been somehow inspired by Bock [38] or more recently by Gordon [39]. A parallel work [15] develops a theory of grouping, but for a completely different application, namely planar shape recognition. For the sake of completeness, the main results of this theory are developed here in the context of motion analysis.

## 4.1 Number of false alarms of a group and cluster validity

The fact that some of the $X_i$'s may be a group reveals a lack of independence of these points. Since the cause of the dependence is unknown, modeling the probability of such an event is difficult. Hence, the idea of the *a contrario* decision is that groups can be detected as large deviations from an independence model. Let us introduce the following *background model*.

**Definition 1** *A* background model *is defined as a stochastic process* $(X_1, \ldots, X_M)$ *which components are independent and identically distributed (i.i.d.) with distribution* $\pi$.

In other words, the *background model* hypothesizes a random organization of the observations. This setting is generic. The case specific probability distribution $\pi$ will be specified later.

Let $R \in \mathbb{R}^4$ be a region independent of the $X_i$'s. Under the hypotheses of the *background model*, the probability that at least $k$ out of the $M$ data points $\{X_1, ..., X_M\}$ belong to $R$ is given by the tail of a binomial law with parameters $k$, $M$, and $\pi(R)$

$$B(M, k, \pi(R)) = \sum_{j=k}^{M} \binom{M}{j} \pi(R)^j (1 - \pi(R))^{M-j}. \tag{1}$$

Let us assume that such a region $R$ containing $k$ data points is observed. If the above probability happens to be very low, the observed data points certainly contradict the i.i.d. hypothesis. Of course, $R$ must be given before observing the data points. From now on, an *a priori* finite set of regions $\mathcal{R}$ with cardinality $|\mathcal{R}|$ is considered, typically hyper-rectangles, centered on the origin.

Let us introduce the following measure of meaningfulness.

**Definition 2** *Let* $G$ *be a subset of* $\{X_1, ..., X_M\}$ *of cardinality* $k$, $2 \leqslant k \leqslant M$. *The number of False Alarms (NFA) of a group* $G$ *is defined as*

$$NFA(G) = M^2 \cdot |\mathcal{R}| \min_{\substack{x \in G, R \in \mathcal{R} \\ G \subset x + R}} B(M - 1, k - 1, \pi(x + R)). \tag{2}$$

*A group* $G$ *is said to be* $\varepsilon$-meaningful *if* $NFA(G) \leqslant \varepsilon$.

Before giving a mathematical result explaining why this number is introduced, let us explain how it is computed. Let us examine the term which appears in the minimum function: $x + R$ is one of the possible regions of $\mathcal{R}$, after centering at $x$, which is a point of $G$. Hence, $B(M - 1, k - 1, \pi(x + R))$ is the probability that at least $k$ points (including $x$) are inside $x + R$ under the hypotheses of the *background model.* Then, $x$ and $R$ are chosen to minimize this probability. Let us remark that there are at most $M|\mathcal{R}|$ possible choices of the couple $(x, R)$. The second factor $M$ is explained in the following.

Let us also give a qualitative explanation of this definition. Up to a multiplicative constant, the NFA measures the probability according to the background model that all the points of $G$ belong to a region centered at a point which is also in $G$. The lower the NFA, the stronger the contradiction to the *background model* and the more meaningful the group. The quantitative meaning is given by Proposition 1 in the following result section.

## 4.2 A set of candidate groups

There are $2^M$ subsets of $\{X_1, ..., X_M\}$. It is not possible to compute the NFA of every possible group. Most of them are anyway certainly irrelevant. In order to drastically reduce the number of candidate groups, a single linkage hierarchical clustering procedure is applied [37]. The result is a binary inclusion tree. Each node of the tree is a candidate group. The root of the tree contains all the $M$ data points. Other clustering algorithms proposing a reasonable set of candidate groups could be considered. Single linkage hierarchical clustering was adopted because it is well suited for processing elongated clusters. The hierarchical structure of the binary tree of candidate groups is useful when dealing with cluster merging issues as explained in the next section.

**Remark** The hierarchical clustering step does not solve the two problems at hand: number of clusters and meaningfulness or validity of each cluster. It only proposes a hierarchy of partitions of the data set. From this procedure, $M-1$ candidate groups containing more than 2 points are proposed. It is then possible to prove the following result.

**Proposition 1** *If $X_1$, ..., $X_M$ are i.i.d. points from the distribution $\pi$, then the expectation of the number of $\varepsilon$-*meaningful *groups among any set of $M$ candidate groups is less than $\varepsilon$.*

In particular, the result holds for the set of candidate groups provided by the hierarchical clustering procedure, since there are less than $M$ candidates. We refer the reader to [15] for a complete proof but let us give a short sketch. A group $G$ is *$\varepsilon$-meaningful*, if there is a couple $(x, R)$ such that $G \subset x + R$ and $B(M-1, k-1, \pi(x+R)) \leqslant \frac{\varepsilon}{M^2|\mathcal{R}|}$. Because the number of points in a given region follows a binomial law, easy (but careful) calculations show that the probability of the above event is less than $\frac{\varepsilon}{M^2|\mathcal{R}|}$. Now, for each candidate group, at most $M|\mathcal{R}|$ possible configurations (corresponding to the choices of $x$ and $R$) are tested. Since at most $M$ groups are tested, the result follows by additivity of the expectation.

The interpretation of this result is more important than its proof. Set $\varepsilon$ to a small value, less than 1. If an *$\varepsilon$-meaningful* group is observed, then chance alone is certainly not a good explanation for it, since less than $\varepsilon < 1$ such meaningful groups would be observed on average if the data is distributed according to the *background model*. The lower the NFA, the less likely it is that such a group has been generated by the *background model*. Hence, the NFA provides a validity measure. In other words, the NFA is a confidence level directly related to the average number of occurrences of the observed event under the hypotheses of the *background model*: the lower the NFA, the more relevant the observed event, the stronger the confidence in the detection. In general, the NFA of a meaningful group is much lower than 1 (see Sect. 5).

The next important question is to select the right representation of a given set of data points: should it be considered as one large group or two smaller groups?

## 4.3 Merging criterion

How to distinguish two close objects from a single large one including them both? The answer is often semantic, which is out of the scope of this paper. We can think for instance of a car taking over another one, with about the same velocity, or two people walking together. However, if the velocities are different enough, or if the objects are far enough from each other, it should be possible to suitably separate observations into distinct groups. In other terms, a merging criterion is needed. Thus, let us consider two disjoint groups $G_1$ and $G_2$, and a group $G$ such that $G_1 \cup G_2 \subset G$. Following the Helmholtz principle, a single group is preferred if it is more unlikely to occur according to the *background model* than two jointly observed groups. Hence, an NFA must be defined for pairs of groups. Let us denote by

$$\binom{M}{i,j} = \frac{M!}{i!j!(M-i-j)!},$$

the trinomial coefficient. Assume that $R_1$ and $R_2$ are two disjoint regions with respective probability $\pi_1$ and $\pi_2$. The probability that at least $k_1$ points belong to $R_1$ and $k_2$ points belong to $R_2$ is

$$\mathcal{M}(M, k_1, k_2, \pi_1, \pi_2) = \sum_{\substack{i \geqslant k_1 \\ j \geqslant k_2}} \binom{M}{i,j} \pi_1^i \pi_2^j (1 - \pi_1 - \pi_2)^{M-i-j}. \tag{3}$$

**Definition 3** *The number of false alarms of the disjoint pair* $(G_1, G_2)$ *is defined as*

$$\begin{aligned} NFA_g(G_1, G_2) = M^4 |\mathcal{R}|^2 \cdot \\ \min_{x_1, x_2, R_1, R_2} \mathcal{M}(M-2, k_1-1, k_2-1, \pi_1, \pi_2) \end{aligned} \tag{4}$$

(cf. Appendix for technical details and exact definition of $k_1$, $k_2$, $\pi_1$ and $\pi_2$). Using the same kind of arguments as for Prop. 1, one can prove that, on average, there are less than $\varepsilon$ pairs with $NFA_g$ less than $\varepsilon$. More interestingly, the normalization of probabilities into NFAs allows comparisons between events of different nature, such as groups and pairs of groups, because the numbers of false alarms have comparable magnitudes.

**Definition 4** *Let $G$ be a subset of the $M$ data points. A group $G$ is said* indivisible*, if and only if, for all pairs $G_1$ and $G_2$ such that $G_1 \cap G_2 = \emptyset$ and $G_1 \cup G_2 \subset G$,*

$$NFA(G) < NFA_g(G_1, G_2).$$

The hierarchy provided by the tree of candidate groups allows us to simplify the problem of deciding to merge two small groups into a larger one. Indeed, since the tree of clusters is binary, this question can be answered for two sibling nodes. The merging method is then applied recursively.

## 4.4 Practical algorithm

So far, a group validity criterion and a merging criterion have been defined. A group is valid if its NFA is less than $\varepsilon = 1$. It should not be split into two smaller groups if it is indivisible. The last point is that a group can be slightly enlarged by adding a few points. Again, does this result in a better representation of the data ? This question is easily answered by comparing the NFAs of the groups through the binary inclusion tree provided by the hierarchical clustering step.

**Definition 5** *A group $G$ is said to be* maximal $\varepsilon$-meaningful *if*

1. *$G$ is* $\varepsilon$-meaningful
2. *$G$ is indivisible.*
3. *$G$ is more meaningful than all its indivisible child nodes.*
4. *for all indivisible parent nodes $G'$, either $NFA(G) < NFA(G')$ or there exists another indivisible child node $G''$ of $G'$ such that $NFA(G'') < NFA(G')$.*

The last condition only reflects that the tree is an asymmetric graph and ensures that a group can eliminate smaller groups (child nodes) in the tree only if it is more meaningful than *all* of them.

All these definitions may seem a bit formal. Actually, the implementation basically reduces to counting points in hyper-rectangles. Let us sum up the meaningful group detection algorithm.

1. **Clustering step**. Given $M$ data points, compute the binary tree of candidate groups by a hierarchical single linkage clustering algorithm. Each node corresponds to a candidate group.
2. **Validity step**. For each candidate group $G$,
   (a) compute the region $x + R$, $x \in G$, $R \in \mathcal{R}$ containing all the points of $G$ and such that $\pi(x + R)$ is minimal.
   (b) compute $NFA(G)$ and tag $G$ as valid if $NFA(G) \leqslant \varepsilon$.
3. **Merging step**. For each sibling pair $G_1$ and $G_2$.
   (a) Compute the intersection of $x_1 + R_1$ and $x_2 + R_2$, obtained in the computation of $NFA(G_1)$ and $NFA(G_2)$.
   (b) Remove the points of $G_1$ and $G_2$ in this intersection.
   (c) Compute $NFA_g(G_1, G_2)$.
4. **Final step**. Explore the tree and detect maximal meaningful groups according to Def. 5.

The last details to be specified are the choice of the *a priori* distribution $\pi$ and the set of regions $\mathcal{R}$. Although the grouping method described so far is generic, the choice of $\pi$ is more problem-specific. In the case at hand, position and velocity of objects are considered independent. Of course this is not true for real objects (for instance, vehicles hopefully follow tracks!). However, the *a contrario* hypotheses describe the absence of correlation of all the observations. Hence, it is sound to assume that the velocity and the position are independent. Moreover, unless it has been specified by the application, the position of a moving object is arbitrary. Hence, the position distribution is assumed to be uniform. No direction plays a particular role either. Hence, the velocity direction distribution is chosen uniform in $(0°, 360°)$. The only problem is the norm of the velocity. Without prior knowledge, specifying a distribution for the velocity magnitude is not obvious. A simple solution is to learn it on the data itself: the distribution of the velocity magnitudes is given by the empirical histogram of the observed velocity magnitudes. This provides the right order of magnitude and a fair enough distribution profile. Now, the joint distribution $\pi$ of the data points is simply the product of these four marginal distributions.

Since the four dimensions of the grouping space $(x^{\text{ref}}, v)$ are assumed uncorrelated, it does make sense to consider regions which main directions are parallel to the axes of coordinates. Moreover, the clusters that have to be found do not have any particular shapes. This results in a set $\mathcal{R}$ of hyper-rectangles with quantized size in each dimension. The set of regions $\mathcal{R}$ is defined as a set of hyper-rectangles, which sizes in each dimension are the terms of a geometric progression of the type $a_0 r^k$, for some fixed $a_0$ and $r > 1$. If $k$ is constrained to $0 \leqslant k \leqslant K$ and the data belongs to $\mathbb{R}^N$, then $|\mathcal{R}| = K^N$. In practice, $K = 20$. Its precise value does not have a large influence on detection results and $K$ is not a sensitive parameter. The value of $a_0$ depends on the accuracy of the considered dimension (position, velocity magnitude, velocity orientation). Therefore, $a_0 = 1$ pixel for initial position, $a_0 = 1$ pixel/frame for velocity magnitude and 0.5 degrees for velocity orientation. Again, the specific values of $a_0$ do not have a strong influence on the detection results as long as the proposed set of regions $\mathcal{R}$ reasonably describes the grouping space $(x^{\text{ref}}, v)$.

# 5 Experimental results

This section presents results for the proposed coherent motion detection method applied to various image sequences. The first experiment aims at checking the validity of the background model and the robustness to false alarms. The second set of experiments illustrates the grouping of local motion measurements obtained by matching the more descriptive features : Similarity Invariant Pieces of Level Lines (SIPLL) and SIFT descriptors. The cluster detection algorithm is applied once to all the local displacement measurements and once to *moving features* only (cf. Section 3), without significative differences in the results. The third set of experiments relies on local displacement measurements computed with the KLT technique. The last experiment shows how the method enables to group displacement measurements corresponding to moving objects undergoing occlusion.

## 5.1 Checking the *background model*

In order to check the relevance of the specified *a contrario* model, the first experiment involves an image sequence in which local motion measurements display no coherence. This independence of the local motion measurements agrees with the specified *background model*. As a consequence, no coherent motion should be detected. This first experiment also intends to test the robustness to false alarms of the method. Therefore, the first video sequence (see Fig. 3) corresponds to a moving water texture. The image sequence consists of 100 frames processed as 10 segments of 10 frames.

SIFT descriptors are matched in successive frames as explained in Section 3 to obtain local displacement measurements. Local motion measurements are accumulated over 10 frames. The corresponding data points in the motion grouping space are displayed in the lower row of Fig. 3. The automatic cluster detection procedure does not detect any group : no false alarm is raised. This agrees with the theory: with $\epsilon$ set to 1, less than one false alarm is observed on average if the measurements are distributed according to the *background model*.

## 5.2 Experiments with similarity invariant pieces of level lines and SIFT descriptors

After checking the robustness to false alarms in the absence of moving objects, let us discuss the results of the independent motion pattern detection method on several image sequences containing up to seven moving objects. As explained in Section 3, the proposed group detection method can be applied to either all local motion measurements or only to a subset, termed *moving features*, after discarding measurements obviously belonging to the static background. Results are qualitatively equivalent while the computational burden of the cluster detection task is lightened.

### 5.2.1 Highway sequence

The independent motion pattern detection method was applied to the highway sequence displayed in Fig. 2 using two types of features presented in Section 3, namely SIFT descriptors and SIPLL. Results of the clustering procedure with these different inputs are shown in Fig. 4 and Fig. 5. The three moving objects that appear in the sequence are detected using either type of features. SIPLL perform better in describing small shapes. When restricting the clustering procedure to the *moving SIPLL*, even the small car in the background of the left lane is detected. This is not possible when clustering all SIPLL, including the static background, since the speed of this car is too low. It is therefore merged with the static background. The quantity $-\log_{10}(NFA)$ measures the confidence in each detected cluster. It increases with the quantity (number of points) and quality (density) of the evidence for each coherent motion. The elongated shape in the velocity magnitude dimension of the cluster corresponding to the car that is the closest to the camera reflects the variation of the velocity of the projections of the object points on the image plane. Taking into account the scene geometry would enable to obtain more concentrated clusters.

### 5.2.2 Parking lot sequence

The second experiment shows rigid and non-rigid moving objects, respectively a car and a pedestrian. Local motion measurements are extracted using SIFT descriptors. Again, a structured description of the dynamic content of the scene is correctly recovered. Here, the confidence levels reflect the nature of the moving objects. The cluster corresponding to the car which is a large rigid object has a confidence level $-\log_{10}(NFA)$ close to 100. This high confidence value is due to the large number of points in the cluster and a high accuracy of the velocity direction. The cluster corresponding to the smaller non-rigid moving pedestrian contains less points. Moreover, their corresponding directions are less steady. Therefore, the confidence level is only about 10. Let us point out that the trees in the background of the scene are moving because of a strong wind. This motion is correctly not detected as coherent when applying the clustering to all the features of the scene.

### 5.2.3 Street sequence

The third processed sequence is again a typical video surveillance sequence (Fig. 7). Several moving objects are present in the scene. From foreground to background: a cyclist is moving down the left lane, a group of pedestrians is crossing the road from right to left on the lower crosswalk, a car enters the scene from the left, a pedestrian is moving up on the right crosswalk, another pedestrian is crossing the street to the left on the upper crosswalk. Finally at a distance, a pedestrian is moving up on the left sidewalk and a car is moving down in the left lane. This seven moving objects are correctly picked out when applying the clustering procedure to the *moving SIFT features* extracted from 20 successive frames. The associated confidence levels $-\log(NFA)$ range from 2 to 50 depending on the size of the moving objects and the characteristics of their motion (magnitude of displacement, steadiness of direction). Small moving objects (single pedestrians) or objects that appear only in the first frames (the cyclist exits the scene at frame 10) have a low confidence level between 2 and 5. The group of pedestrians on the lower crosswalk has a confidence level of 11 and the car has a confidence level of 50. SIFT descriptors are preferred because of the low quality of the image sequence. Strong MPEG compression causes the contours of shapes to be unstable. This perturbs the matching of SIPLL. For each cluster, the mean velocity can be computed. This quantity is a good estimation of the motion of the objects as illustrated by the first row of Fig. 7. When applying the mean velocity computed for each cluster to the region defined in the reference frame (here the first frame), this region precisely follows the position of the moving objects in the successive frames.

Computation time greatly depends on the number of features involved which usually increases with the size of the image. As an example, for the sequence corresponding to Fig. 4 (10 frames of size $352 \times 288$), it takes about 3 seconds to extract the *moving SIFT descriptors* and to cluster the 51 pairs of features. Extracting all the SIFT descriptors and clustering takes about 20 seconds.

Discarding features corresponding to the static background of the scene decreases the computational cost of the clustering procedure through the reduction of the number of considered observations. It has almost no influence on the performance of the method but greatly simplifies the grouping task.

The next subsection presents results obtained with the KLT features and a direct computation of displacements instead of resorting to correspondences between features.

## 5.3 Experiments with *moving KLT features*

In this section, experiments carried out using KLT features are reported. These features are less descriptive than SIFT descriptors and SIPLL. However, the simplicity and the low computational cost of KLT features are very attractive. In order to explicit this simplicity, a different way of discarding background features is preferred : features with a displacement lower than 1 pixel are eliminated after dominant motion compensation. This threshold is chosen in order to agree with the spatial image sampling rate and is not very demanding. Let us stress again that the aim of this step is to remove displacements belonging to the static background in order to lighten the burden of the clustering algorithm. Among all the measurements with sufficient displacement magnitude (larger than 1 pixel/frame), the coherent motion detection algorithm groups those belonging to the same moving objects and discards spurious motion measurements as outliers. The resulting algorithm is fast and self-contained.

### 5.3.1 Coastguard sequence

The experiment on the "Coastguard" sequence (Fig. 8) demonstrates the efficiency of our coherent motion detection method. Two ships are crossing each other. The camera is tracking first the smaller ship and then the larger one. The detection of coherent motion on this sequence is rather challenging due to the presence of a moving texture (water). Instantaneous motion detection techniques (based on 2 or 3 frames, without any prior knowledge on the scene content) should detect water movements as moving regions.

The proposed coherent motion detection method is applied on temporal segments of five successive frames. The method successfully groups the local motion measurements belonging to each ship while observations corresponding to water are discarded as outliers. On this short time interval (five frames), sufficient evidence is gathered in favor of coherent motion in order to detect only motions displaying persistent characteristics over time. The associated NFAs are already extremely low meaning that the confidence in the detection is high.

### 5.3.2 Pedestrian sequence

The next sequence (Fig. 9 and Fig. 10) contains a pedestrian walking on a sidewalk and illustrates the behavior of the detection method on articulated objects. The camera is tracking the pedestrian. The tree and the bushes in the foreground are moving due to the wind. Local motion measurements are accumulated over 10 frames. The cluster detection is applied only to *moving features* (cf. Section 3).

In the first part of the video, the unoccluded torso of the pedestrian is detected as a moving regions. The NFA is very low and thus the confidence in the detection is very high ($-\log_{10}(NFA) = 168$). In the second part of the sequence, the pedestrian is partially occluded by the branches of the tree. Only a few motion measurements are still available. However, the pedestrian is still detected. Of course, the confidence in the detection is then lower ($-\log_{10}(NFA) = 24$) reflecting the fact that there is less evidence in favour of coherent motion.

## 5.4 Moving objects undergoing occlusions

The last part of this experimental section is concerned with moving objects undergoing occlusion. The proposed coherent motion detection method succeeds in grouping together local motion measurements before and after occlusion. The number of frames involved in this experiments is larger (15-30 frames) in order to observe the objects before and after occlusion.

### 5.4.1 Car sequence

This first sequence (Fig. 11) shows a car passing behind a map sign. The camera is shaking while tracking the car. The coherent motion detection procedure is applied to 30 frames. In the first frame, only the back of the car is visible. In the last, only the front part of the car appears in the image. The car is never visible in its whole. Based on the proximity in the velocity space, all the measurements corresponding to the car are clustered together. Due to the large number of measurements inside the cluster (451 points), caused by the validity of the constant motion hypothesis and the long observation duration (30 frames), the confidence in this detection is extremely high, $-log_{10}(NFA) = 308$.

### 5.4.2 Crossing pedestrian sequence

The second sequence, Fig. 12, is slightly more complex: a pedestrian is crossing another one and gets occluded. The camera is hand held and is tracking the first pedestrian. Local motion measurements are accumulated through 15 successive frames. Both pedestrians are detected as undergoing coherent motions. The local motion measurements belonging to each of them are clustered into two separate groups. Outliers correspond to measurements due to noise or measurements on the arms and legs having a periodic motion that does not display sufficient coherence.

## 5.5 Number of frames involved in the detection process

The number of frames during which motion information is accumulated can vary. Part of this work was to study how long an image sequence has to be examined in order to detect groups of coherent motion. The conclusion is that several factors influence the minimal observation time : size of objects, image quality, validity of the first order approximation on the trajectory (constant velocity). It turns out that under favorable conditions, the required number of frames can be as small as 3 or 5. The number of frames involved in the coherent motion detection process can be tuned according to the specific application and the experimental conditions:

- 3-5 frames: "instantaneous" motion detection enforcing motion coherence;
- 5-10 frames: short-term coherent motion detection;
- 10-30 frames: long-term coherent motion detection, especially in the case of occlusions.

Let us point out that in all cases the observation time remains short: less than one second for 30 frames/second video.

# 6 Conclusions and perspectives

This paper presents a method to detect independent coherent motion patterns in image sequences. The automatic clustering of local motion measurements leads to a general *coherent motion* detection algorithm. The result is a structured description of the dynamic scene content: number of moving objects, position, magnitude and direction of their displacements, i.e., local trajectories. The proposed framework enables to control the number of false alarms and associates a confidence level to each detected independent motion pattern. The local motion measurements are extracted by means of characteristic image features. Possible types of image features are: similarity invariant pieces of level lines, SIFT descriptors or KLT features. Results on various real image sequences illustrate the ability of the method to detect temporally consistently moving objects (cars, pedestrians) without being distracted by moving textures (water, leaves). Future work will aim at extending the proposed method to 3D motion models. If the scene geometry is known it could be incorporated into the model to take into account variation of the projected velocity due to depth changes of moving objects. As for the local trajectories provided as an output of the method, they could become useful for long term trajectory analysis. Further work on the clustering algorithm itself consists in processing partial trees in order to reduce computation time. The description of the scene provided by this method could become useful for surveillance, activity recognition, as well as robot navigation.

# Appendix

The rigorous definition of an NFA for a pair of disjoint groups (Def. 3) requires the following technical precautions. Let $G_1$ and $G_2$ be two disjoint sets of points and consider two regions of the type $x_1 + R_1$ and $x_2 + R_2$ centered at $x_1 \in G_1$ and $x_2 \in G_2$. Although the groups are disjoint, these regions may intersect. Let

$$k_1 = |G_1 \backslash (x_2 + R_2)| \text{ and } k_2 = |G_2 \backslash (x_1 + R_1)|$$

be the number of points of $G_1$ (resp. $G_2$) that are not in $x_2 + R_2$ (resp. $x_1 + R_1$). Let also

$$\pi_1 = \pi((x_1 + R_1) \backslash (x_2 + R_2))$$

be the probability that one point belongs to $x_1 + R_1$ while avoiding $x_2 + R_2$. Symmetrically, let $\pi_2 = \pi((x_2 + R_2) \backslash (x_1 + R_1))$. The NFA of the disjoint pair $(G_1, G_2)$ is then given by Eq. (4).

In practice, the following simplifications are applied. First, in order to compute $NFA_g(G_1, G_2)$ (Def. 4), the regions $x_1 + R_1$ and $x_2 + R_2$ that have already been obtained in the computation of $NFA(G_1)$ and $NFA(G_2)$ are considered. Moreover, not all the possible pairs of disjoint subsets of $G$ are tested but only those appearing in the tree of candidate groups provided by the hierarchical clustering step.

# References


[1] S. Ullman, The Interpretation of Visual Motion, 2nd Edition, MIT Press, 1982.

[2] Y. Hsu, H.H.Nagel, G. Rekers, New likelihood test methods for change detection in image sequences, Computer Vision, Graphics and Image Processing 26 (1984) 73–106.

[3] P. Rosin, Thresholding for change detection, Computer Vision and Image Understanding 86 (2002) 79–95.

[4] T. Aach, A. Kaup, Bayesian algorithms for change detection in image sequences using Markov random fields, Signal Processing: Image Communication 7 (2) (1995) 147–160.

[5] P. Pérez, C. Hue, J. Vermaak, M. Gangnet, Color-based probabilistic tracking, in: Eur. Conf. on Computer Vision, ECCV'2002, LNCS 2350, Copenhagen, Denmark, 2002, pp. 661–675.

[6] E. Arnaud, E. Mémin, B. Cernuschi-Frias, Conditional filters for image sequence based tracking - application to point tracking, IEEE Trans. on Image Processing 14 (1) (2005) 63–79.

[7] R. Venkatesh Babu, P. Pérez, P. Bouthemy, Kernel-based robust tracking for objects undergoing occlusion, in: Proc. Asian Conf. on Comp. Vision (ACCV'06), Hyderabad, India, 2006.

[8] A. Desolneux, L. Moisan, J. Morel, A grouping principle and four applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (4) (2003) 508–513.

[9] F. Attneave, Some informational aspects of visual perception, Psychological Review 61 (3) (1954) 183–193.

[10] D. Lowe, Perceptual Organization and Visual Recognition, Kluwer Academic Publishers, Boston, Mass., 1985.

[11] J. Lisani, L. Moisan, P. Monasse, J. Morel, On the theory of planar shape, SIAM Multiscale Modeling and Simulation 1 (1) (2003) 1–24.

[12] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[13] J. Shi, C. Tomasi, Good features to track, in: IEEE Int. Conf. on Computer Vision and Pattern Recognition, Seattle, 1994, pp. 593–600.

[14] C. Tomasi, T. Kanade, Detection and tracking of point features, Tech. Rep. CMU-CS-91-132, Carnegie Mellon University (April 1991).

[15] F. Cao, J. Delon, A. Desolneux, P. Musé, F. Sur, A unified framework for detecting groups and application to shape recognition, Journal of Mathematical Imaging and Vision, published online.

[16] L. Wixson, Detecting salient motion by accumulating directionally-consistent flow, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 774–780.

[17] J. Gryn, R. Wildes, J. Tsotsos, Detecting motion patterns via direction maps with applications to surveillance, in: 7th IEEE workshop on applications of computer vision, 2005, pp. 202–209.

[18] Y. Ricquebourg, P. Bouthemy, Real-time tracking of moving persons by exploiting spatio-temporal image slices, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 797–808.

[19] S. Sarkar, D. Majchrzak, K. Korimilli, Perceptual organization based computational model for robust segmentation of moving objects, Computer Vision and Image Understanding 86 (2002) 141–170.

[20] P. Kornprobst, G. Medioni, Tracking segmented objects using tensor voting, in: International Conference on Computer Vision and Pattern Recognition, Vol. 2, Hilton Head Island, South Carolina, 2000, pp. 118–125.

[21] I. Laptev, S. Belongie, P. Pérez, J. Wills, Periodic motion detection and segmentation via approximate sequence alignment, in: Proc. Int. Conf. on Computer Vision (ICCV'05), Beijing, China, 2005, pp. 816–823.

[22] A. Yuille, N. Grzywacz, A theoretical framework for visual motion, in: T. Watanabe (Ed.), High-Level Motion Processing, MIT Press, 1998.

[23] P. Burgi, A. Yuille, N. Grzywacz, Probabilistic motion estimation based on temporal coherence, Neural Computation 12 (2000) 1839–1867.

[24] Q. Gao, Y. Zhang, A. Parslow, The influence of perceptual grouping on motion detection, Computer Vision and Image Understanding 100 (3) (2005) 442–457.

[25] A. W. Fitzgibbon, A. Zisserman, Multibody structure and motion: 3-D reconstruction of independently moving objects., in: 6th European Conference on Computer Vision, Vol. 1843 of LNCS, Springer, Dublin, 2000, pp. 891–906.

[26] J. P. Costeira, T. Kanade, A multibody factorization method for independently moving objects, Int. J. Comput. Vision 29 (3) (1998) 159–179.

[27] L. Zelnik-Manor, M. Irani, Temporal factorization vs. spatial factorization, in: 8th European Conference on Computer Vision, Vol. 2, Prague, 2004, pp. 434–445.

[28] Y. Lin, Y. Chen, S. Kung, A principal component approach to object-oriented motion segmentation and estimation, Journal of VLSI Signal Processing 17 (1997) 163–187.

[29] R. Hartley, R. Vidal, The multibody trifocal tensor: Motion segmentation from 3 perspective views, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. I, 2004, pp. 769–775.

[30] R. Vidal, S. Sastry, Optimal segmentation of dynamic scenes from two perspective views, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. II, Madison, Wisconsin, 2003, pp. 281–286.

[31] P. H. S. Torr, D. W. Murray, Stochastic motion clustering, in: ECCV '94: Proceedings of the 3rd European Conference on Computer Vision, Vol. 2 of LNCS, Springer, Stockholm, 1994, pp. 328–337.

[32] P. Musé, F. Sur, F. Cao, Y. Gousseau, J. Morel, An a contrario decision method for shape element recognition, International Journal of Computer Vision 69 (3) (2006) 295–315.

[33] C. Tomasi, T. Kanade, Detection and tracking of point features, Tech. Rep. CMU-CS-91-132, Carnegie Mellon University (1991).

[34] C. Harris, M. Stephens, A combined corner and edge detector, in: 4th Alvey Vision Conference, Manchester, 1988, pp. 189–192.

[35] J. Odobez, P. Bouthemy, Robust multiresolution estimation of parametric motion models, Journal of Visual Communication and Image Representation 6 (4) (1995) 348–365, software available at http://www.irisa.fr/vista/Motion2D.

[36] T. Veit, F. Cao, P. Bouthemy, An a contrario decision framework for region-based motion detection, International Journal of Computer Vision 68 (2) (2006) 163–178.

[37] A. Jain, R. Dubes, Algorithms for clustering data, Advanced Reference Series, Prentice-Hall, 1988.

[38] H. Bock, On some significance tests in cluster analysis, Journal of Classification 2 (1985) 77–108.

[39] A. Gordon, Classification, 2nd Edition, Chapman and Hall, 1999.

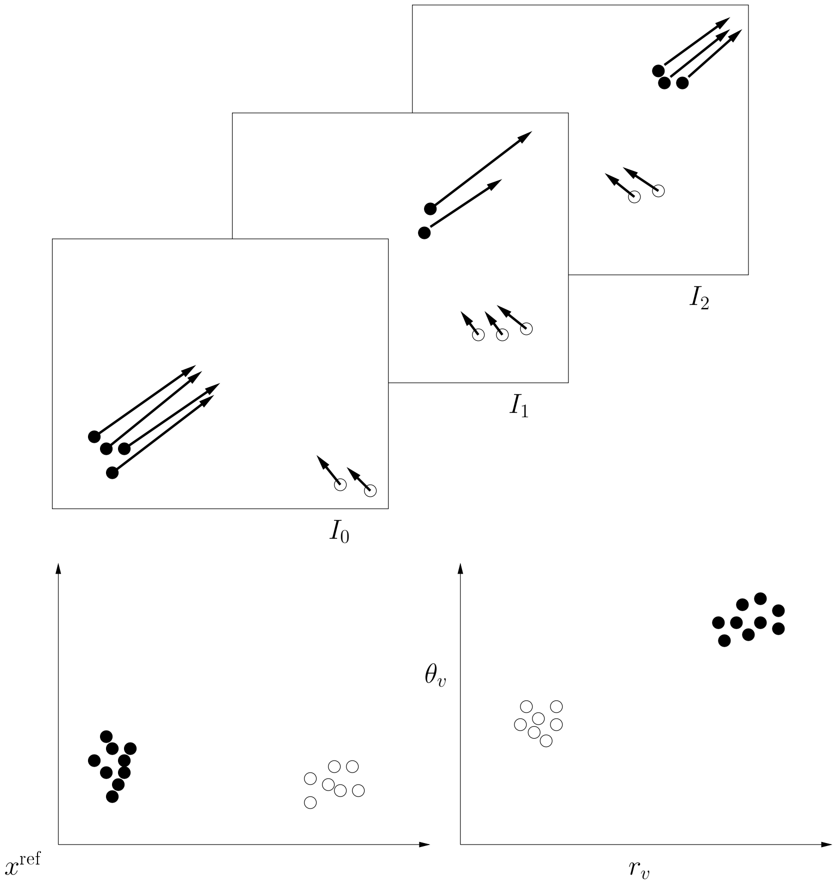


Figure 1: Local motion measurements and corresponding data points in the grouping space. Top row : local motion measurements computed for 3 successive frames. Bottom row : Corresponding data points $(x^{\mathrm{ref}}, v)$ in the grouping space. The velocity $v$ is represented with polar coordinates $v = (r_v, \theta_v)$

.

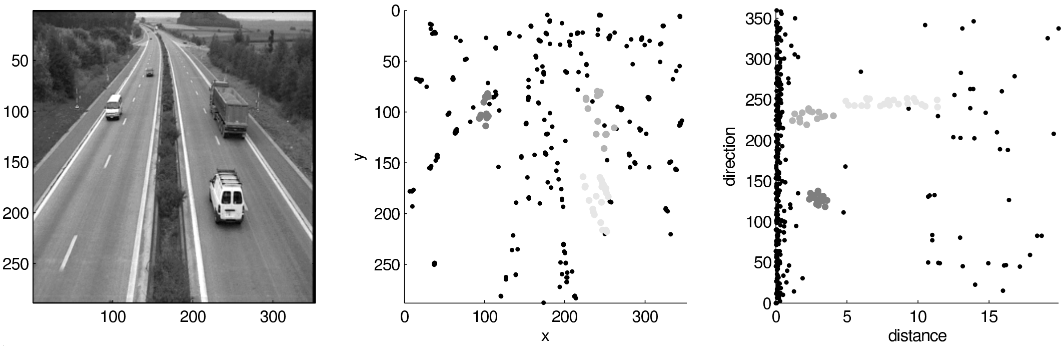


Figure 2: Left image : three moving objects are perceptible, in the left lane a white van, in the right lane a white van and a gray truck. Middle and right plots : two-dimensional projections of four-dimensional couples $(x^{\mathrm{ref}}, v)$, column vs. line of the initial position and velocity direction vs. velocity magnitude. The three moving objects form three distinctive clusters in $\mathbb{R}^4$ (plotted with different grey levels). Elements belonging to the static background appear as a large elongated cluster with almost zero velocity magnitude and no distinct direction.

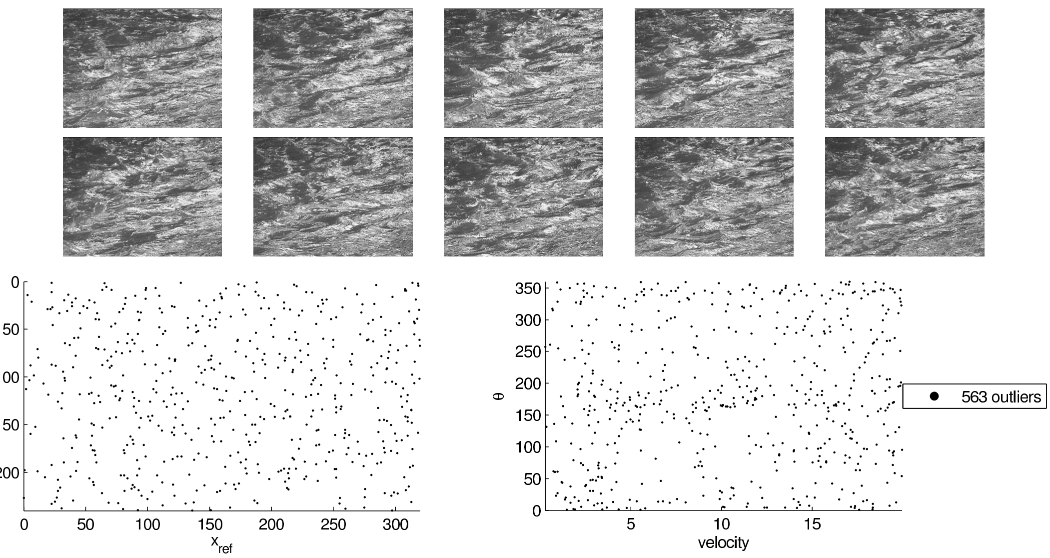


Figure 3: Moving texture of sea-waves. In this first sequence motion measurements are supposed to follow the *background model*. The absence of coherent motion is obvious. The 100 images of the sequences are processed as batches of 10 frames. As expected, no cluster is detected, all data points are classified as outliers.

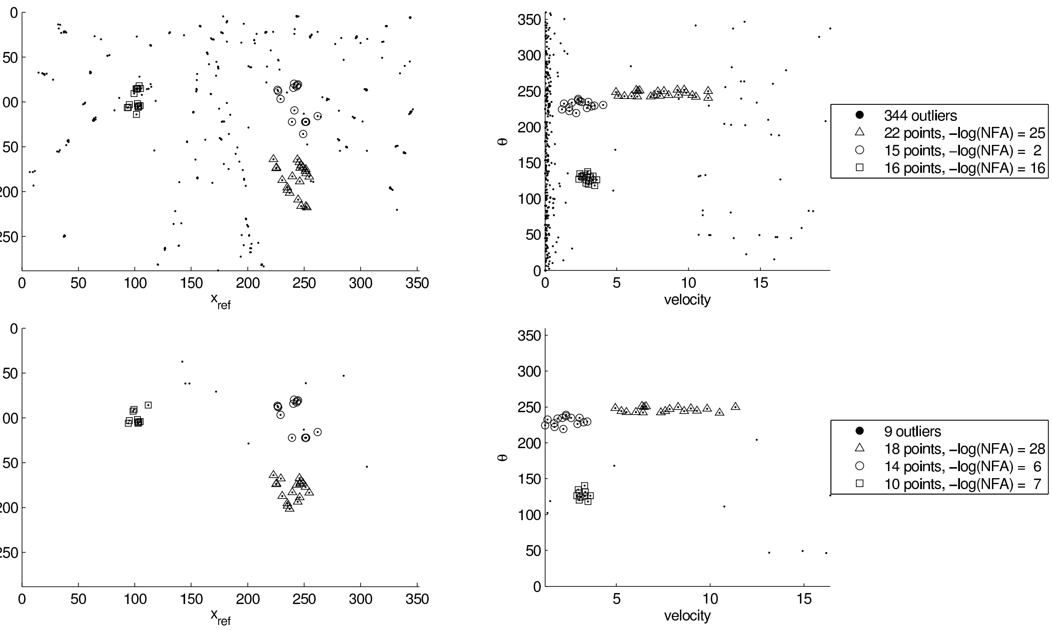


Figure 4: Clustering results on two different sets of data points extracted from 10 frames of the highway sequence (Fig. 2). The four-dimensional motion parameter space is represented by the two two-dimensional subspaces corresponding to initial position (left column) and displacement magnitude and orientation (right column). First row, all SIFT descriptors. Second row, *moving SIFT descriptors* only (cf. Section 3). The confidence levels $-\log(NFA)$ of each detected group appear in the legend on the right. The three moving objects are detected.

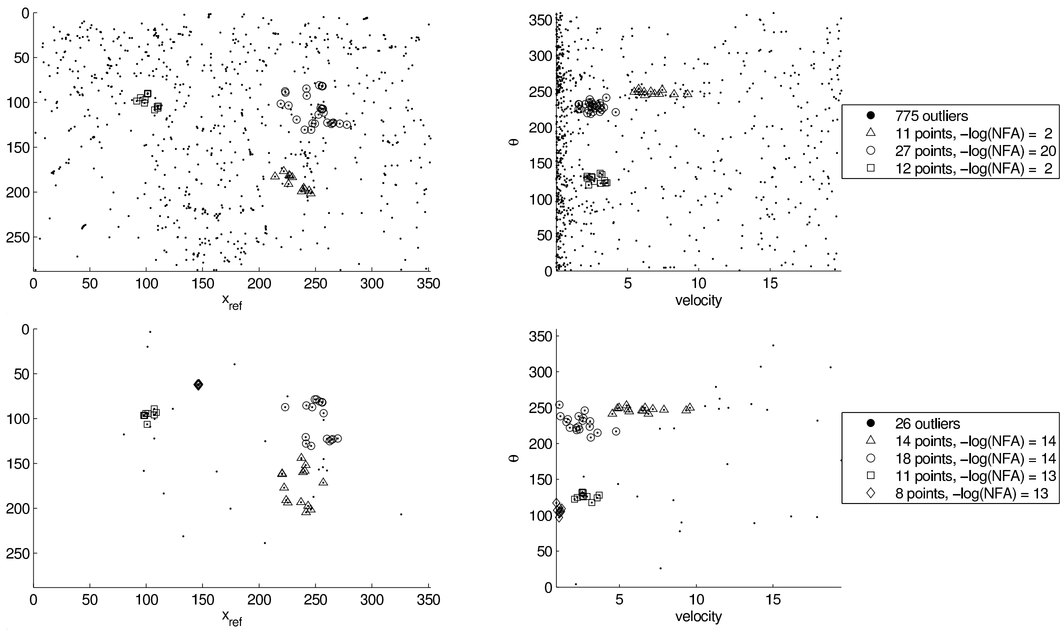


Figure 5: Clustering results on two different sets of data points extracted from 10 frames of the highway sequence (cf. Figure 2). The four-dimensional motion parameter space is represented by the two two-dimensional subspaces corresponding to initial position (left column) and displacement magnitude and orientation (right column). First row, all SIPLL. Second row, *moving SIPLL* only. The clusters corresponding to the three moving objects are detected. Using *moving SIPLL*, even a fourth object is extracted. It corresponds to the car in the background of the left lane which is hardly perceptible and difficult to describe using features. The confidence levels $-\log(NFA)$ of the detected groups appear in the legend on the right.

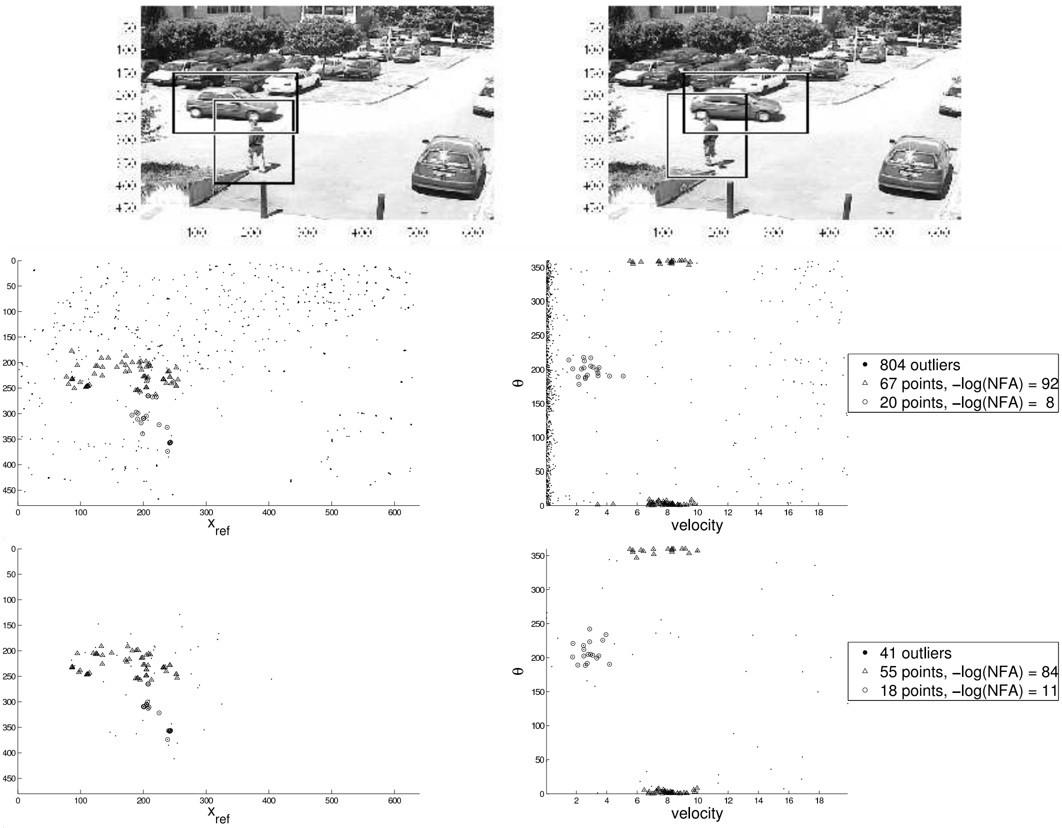


Figure 6: First row: first (t=1) and last (t=10) input frames. In the left image, the black rectangles delineate the regions associated to the clusters when grouping *moving SIFT descriptors*. In the right image, the regions extracted in the reference frame (first frame) are simply moved according to the mean motion of the cluster points. The second row presents the two-dimensional projections of the four-dimensional motion space when considering all the SIFT descriptors of each image. The third row contains the clustering results when considering only *moving SIFT descriptors*. Results with these two options are similar. Clustering only the *moving features* (third row) is of course faster. The confidence levels $-\log(NFA)$ of the detected groups appear in the legend on the right.

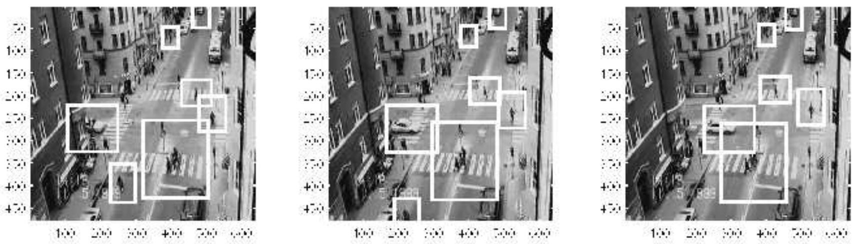

Figure 7: Street sequence : Frame 1, 10 and 20. In frame 1, rectangles correspond the regions associated to each cluster. For frames 10 and 20, the regions are shifted according to the mean motion computed from each cluster. This motion estimation is qualitatively satisfying since the rectangles follow the real motion of the objects. The confidence levels ($-\log(NFA)$) range from 2 for small slowly moving objects to 50 for the car.

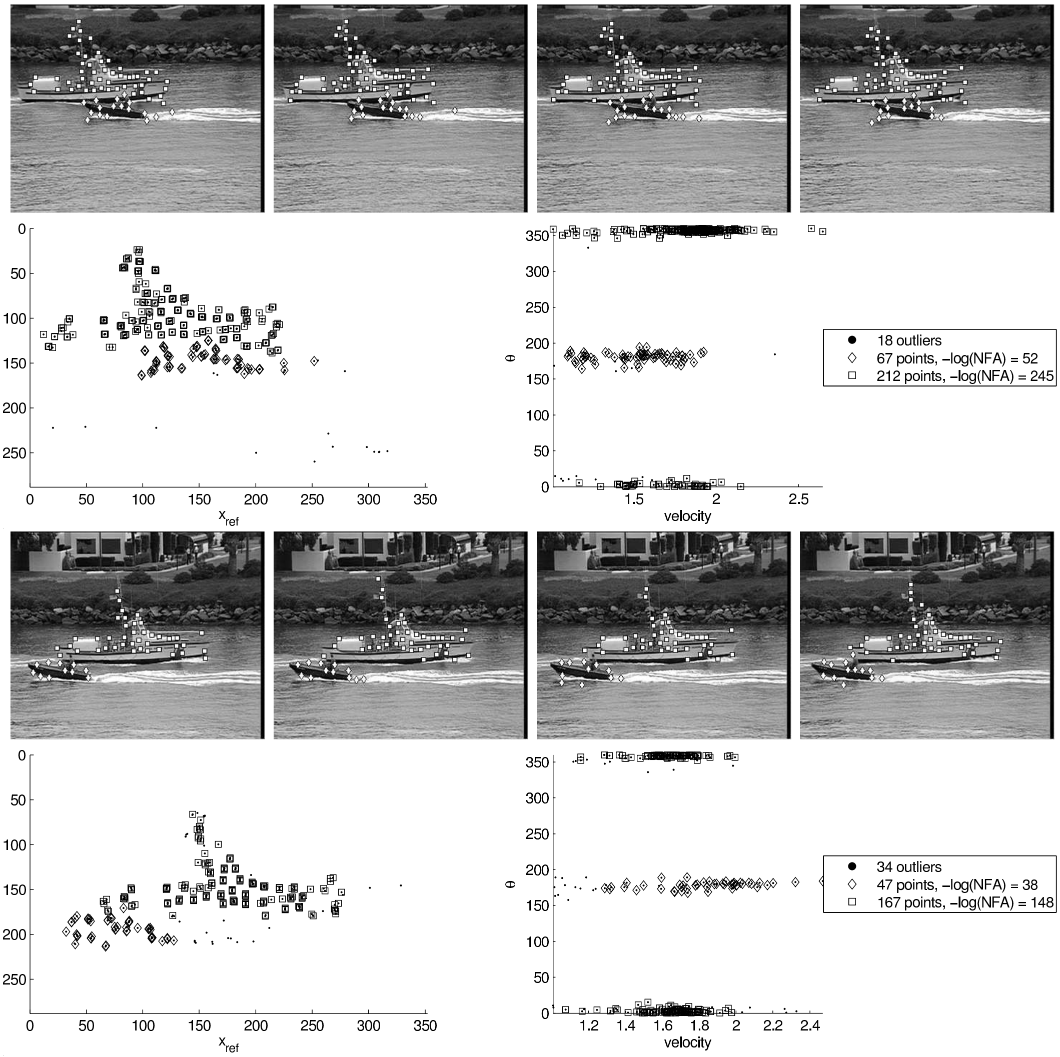


Figure 8: Coastguard sequence. Results on two segments of five frames at two different time instants. Both ships are detected accurately and with high confidence. Local motion measurements on the water and on the wake of the smaller boat are rejected as outliers.

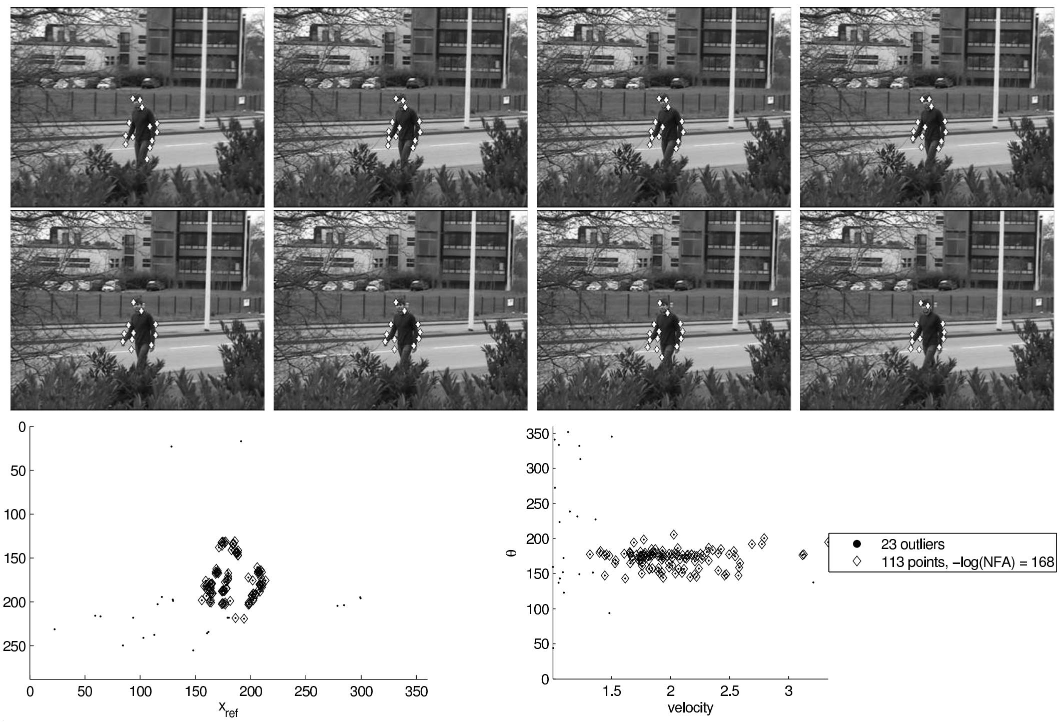


Figure 9: Pedestrian sequence. Local motion measurements are accumulated on 10 successive frames. The pedestrian is detected as a coherent moving region. The oscillating motion of the twigs of the tree and of the bushes is not detected as coherent over time

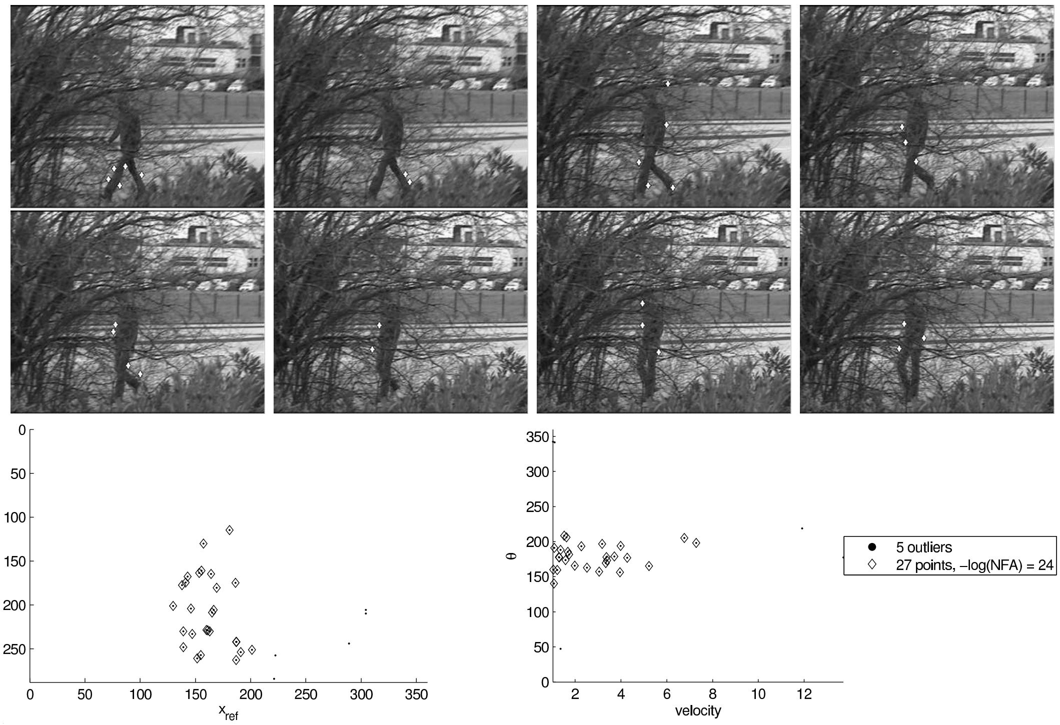


Figure 10: Pedestrian sequence. The pedestrian is now partially occluded. However, sufficient evidence for coherent motion is still available. The confidence in the detection decreases from $-log_{10}(NFA) = 168$ without occlusion to $-log_{10}(NFA) = 24$ when the pedestrian is partially occluded by twigs.

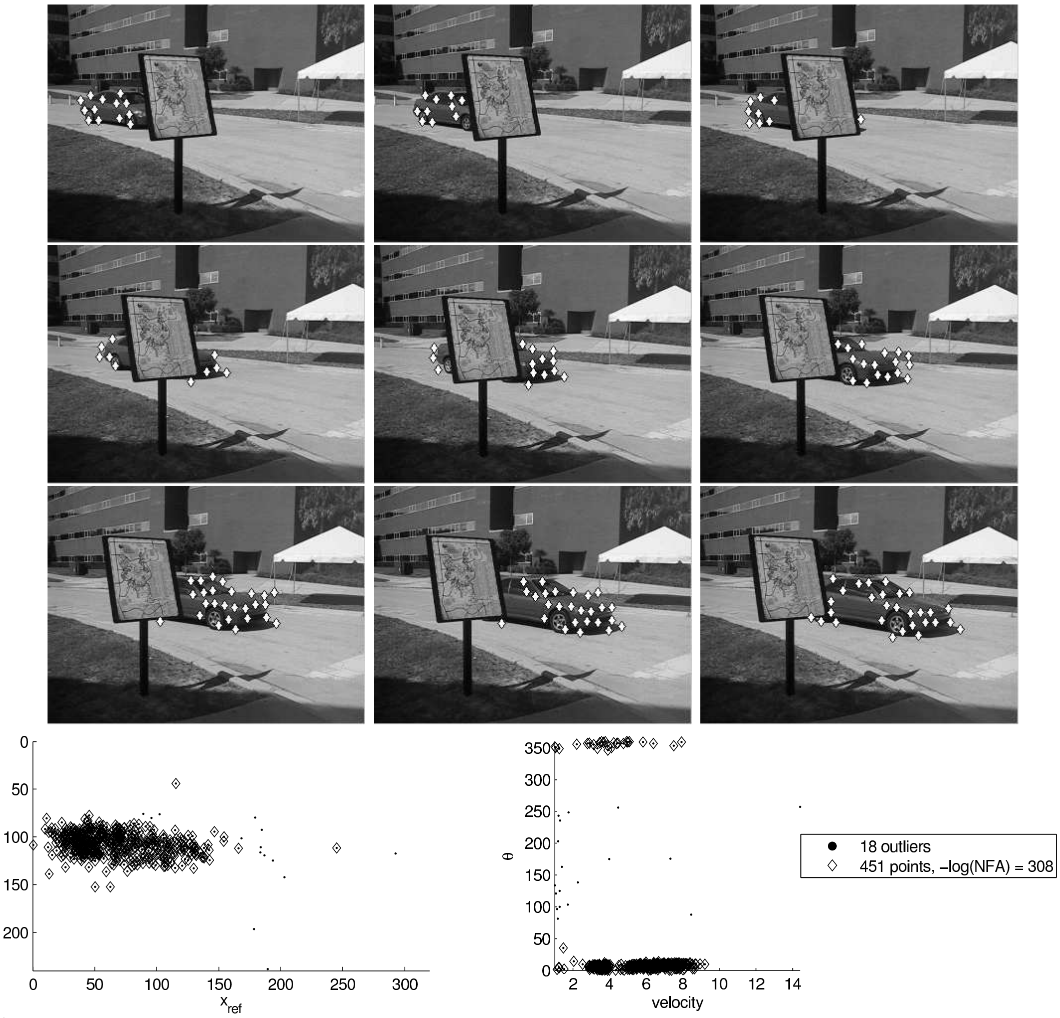


Figure 11: Car sequence. The car is occluded by the signpost. The car is never visible in its whole. Local motion measurements are computed from 30 successive frames. Our coherent motion detection method enables to group motion measurements corresponding to the front and to the back of the car, thus, obtaining a complete description of the car trajectory.

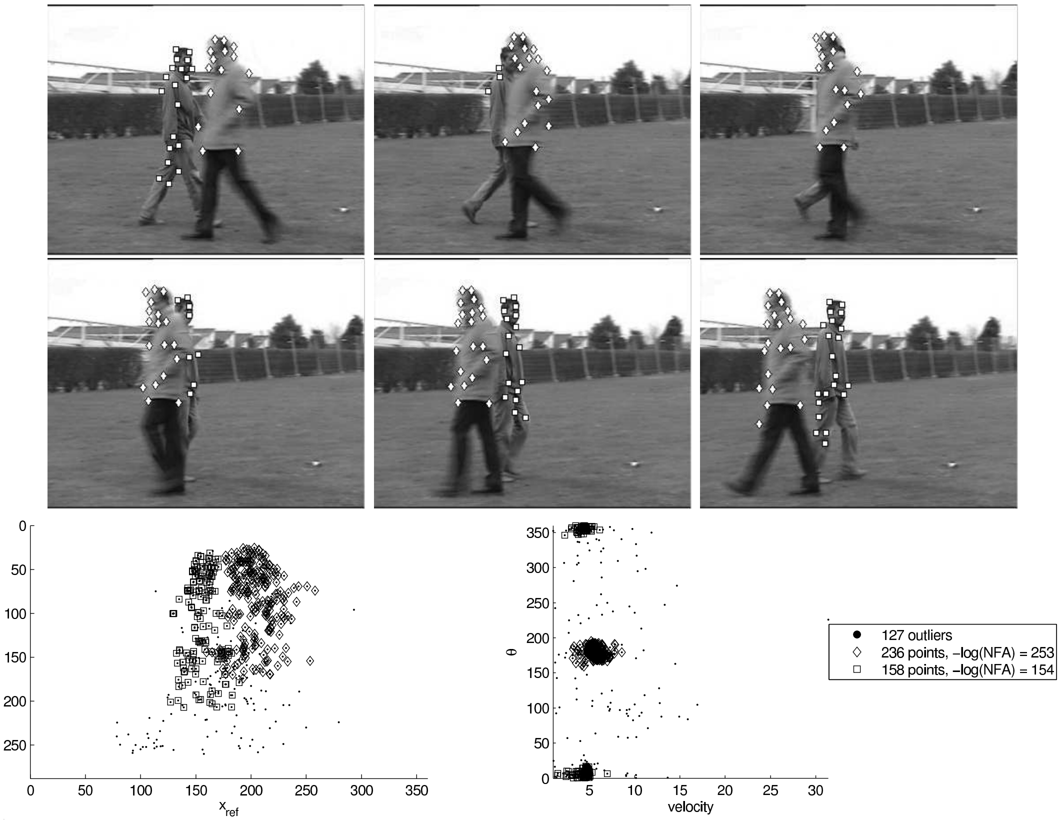


Figure 12: Pedestrians crossing sequence. One pedestrian gets completely occluded by another. The camera is hand held and is tracking the further pedestrian. Local motion measurements are computed from 15 successive frames. Two clusters corresponding to the two pedestrians are detected with very high confidence, $-log_{10}(NFA) = 154$ and $-log_{10}(NFA) = 253$.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399